R Notebook

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

```
## Registered S3 method overwritten by 'Hmisc':
##   method          from
##   summary.formula ergm
```

```
## Registered S3 method overwritten by 'RDS':
##   method         from
##   $.control.list statnet.common
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

Data Exploration

```
bike_data <- readRDS("bike_dat.Rdata") #read data
bike_data <- data.frame(bike_data)
```

```r
str(bike_data)
```

```
## 'data.frame':    856 obs. of  12 variables:
##  $ Date     : Factor w/ 214 levels "1-Apr","1-Aug",..: 1 78 155 173 180 187 194 201 208 8 ...
##  $ Day      : chr  "Friday" "Saturday" "Sunday" "Monday" ...
##  $ High.Temp: num  78.1 55 39.9 44.1 42.1 45 57 46.9 43 48.9 ...
##  $ Low.Temp : num  66 48.9 34 33.1 26.1 30 53.1 44.1 37.9 30.9 ...
##  $ Total    : num  11497 6922 4759 4335 9471 ...
##  $ snow     : chr  "no" "no" "no" "yes" ...
##  $ prec     : num  0.01 0.15 0.09 0.47 0 0 0.09 0.01 0.09 0 ...
##  $ rain     : chr  "yes" "yes" "yes" "no" ...
##  $ day      : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ month    : Factor w/ 7 levels "Apr","Aug","Jul",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ counts   : num  1704 827 526 521 1416 ...
##  $ bridge   : Factor w/ 4 levels "Brooklyn.Bridge",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
sum(is.na(bike_data)) #Check if there are empty values
```

```
## [1] 0
```

```r
train <- bike_data %>%
              sample_n(200, replace = FALSE)
```

```r
#set random seed
set.seed(4168216)

# add year to the date variable
train$Date <- paste(train$Date,"2006", sep="-")
```

```r
#problematic_values <- train$Date[is.na(train$Date)]
#print(problematic_values)
train$Date_col <- as.Date(train$Date, format = "%d-%b-%Y")
```

```r
#Convert categorical variables into factors
train$month <- as.factor(train$month)
train$rain <- as.factor(train$rain)
train$bridge <- as.factor(train$bridge)
#train$snow <- as.factor(train$snow)
train$rain <- as.factor(train$rain)
train$day <- as.factor(train$day)
train$Date_col <- as.factor(train$Date_col)

#levels = c(levels(train$snow))
#print(levels)

print(table(train$snow))
```

```r
#train$snow <- factor(train$snow, levels = c(levels(train$snow), "check"))

#create a linear model
#we remove snow because it has only one value no
mod_lm <- gam(counts ~  s(Date_col, bs="re") + Day + s(High.Temp) + s(Low.Temp) + rain + snow + bridge +




summary(mod_lm)
```

Using a regular linear model is statistically significant as it explains 92.8% of the deviance.

From the data obtained in the summary of mod_lm above we can see that the day of the week that produces the highest number of cyclists is Tuesday as it has the highest estimate at 805.72 and the month of July produces the most cyclists with an estimate of 571.87.

We can see that rain and snow days both have negative estimates (at -315.71 and -1392.73) indicating that a day with rain or snow leads to a lower number of cyclists with much lower number for the latter. We can see that the effects of rain and snow on the number of cyclists are statistically significant as p-values for rainyes = 0.017 and p-values for snowyes = 0.081 are both greater than 0.001.

Since the p-value of High.Temp (9.44e-06) and the p-value of precipitation (1.44x10-5) is much lower than 0.001, it indicates that there is a significant effect of high temperature and precipitation on the number of cyclists. On the other hand, the p-value of Low.Temp is 0.115 which is greater than 0.001, indicating that the effect of low temperature on cyclists is not so significant.

```r
plot(mod_lm, residuals=TRUE, cex=.5, col="blue")  # Interaction plot for var1 and var2
```

From the spline curves above we can see that highest temperature of a day has a complex non-linear relationship with the number of cyclists as moderately high temperatures between 60 to 80 degrees Fahrenheit can actually increase the number of cyclists on the road but extremely high maximum temperatures (those beyond 88 degrees Fahrenheit) result in a decrease in the number of cyclists on the road.

Furthermore, we can also investigate the effects of low temperatures on the number of cyclists. We observe that extremely low temperatures results in a decrease in the number of cyclists. As the lowest temperatures of the day increase, the number of cyclists decrease (indicating that it gets really hot).

We can see that precipitation has an inverse linear relationship (indicated by the small value of edf = 1.001) with the daily count of cyclists as higher precipitation leads to much lower cyclists and vice versa.